DOCUMENT RESUME

ED 450 183 UD 034 004

AUTHOR Madaus, George F.; Clarke, Marguerite
TITLE The Adverse Impact of High Stakes Testing on Minority Students: Evidence from 100 Years of Test Data.
PUB DATE 2001-00-00
NOTE 51p.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Diversity (Student); *Dropout Rate; Elementary Secondary Education; Equal Education; High Risk Students; *High Stakes Tests; *Minority Group Children; Performance Based Assessment; *Student Evaluation; Student Motivation
IDENTIFIERS Adverse Impact

ABSTRACT

This paper examines four aspects of current high stakes testing that impact minority students and others traditionally underserved by American education. Data from research conducted at Boston College over 30 years highlight 4 issues: high stakes, high standards tests do not have a markedly positive effect on teaching and learning; high stakes tests do not motivate the unmotivated; authentic high stakes assessments are not a more equitable way to assess the progress of students who differ in race, culture, native language, or gender; and high stakes testing programs have been shown to increase high school dropout rates, particularly among minority populations. The paper emphasizes the need to carefully monitor the four issues. Though raising educational standards and improving educational quality in American schools is important, efforts to foster academic achievement must involve more than simply setting demanding standards and mandating examinations that are referenced to them. The task remains to identify strategies for achieving the desirable reform objectives efficiently and effectively without having a negative impact on any subpopulation of students. An appendix presents: comparability of the Third International Mathematics and Science Study (TIMSS) Population 3 coverage index and data on dropouts in the United States. (Contains 59 references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 450 183

THE ADVERSE IMPACT OF HIGH STAKES TESTING ON MINORITY STUDENTS: EVIDENCE FROM 100 YEARS OF TEST DATA

George Madaus and Marguerite Clarke

2001

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
X Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

M. Clarke

UD034004

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high stakes testing on minority students: evidence from 100 years of test data. In G. Orfield and M. Kornhaber (Eds.), Raising standards or raising barriers? Inequality and high stakes testing in public education. New York: The Century Foundation.

It's a bull market for high stakes testing programs in education far surpassing the bull market days of minimum competency testing of the early 70s. Now they are called assessments not tests. Their look and feel may have changed but deep down the underlying technology hasn't, and the same issues about their impact and effects persist. The range of high-stakes testing programs is expansive; from "readiness" testing for entrance to kindergarten, to tests required for promotion and graduation, to teacher, school, and district accountability, to teacher testing for certification. These high-stakes testing programs will not go away. If anything they will become more important as policy tools and societal signaling devices. For example, policy makers in several states are setting high standards on state exams to deliver a "wake up call" to students, teachers, parents, and the public to what they portray – incorrectly we would argue-- as a generalized crisis, deterioration, and failure of public education.[1]

The standards movement that is sweeping the country takes for granted that high standards on high-stakes assessments -- in present parlance assessments "worth teaching to" -- can reduce adverse impact in test performance by influencing what is taught and learned. Paradoxically, they are right and wrong. They are right because high-stakes tests do influence what and how things are taught and learned. There is no question about it. The history of high-stakes testing over the last five centuries is testimony to that truism. Test scores will go up. However, as teaching turns into test preparation -- and history again bears witness to this happening -- test results cease to reflect what

examinees really know or can do. They are wrong when they think we can test our way out of our educational problems; the opposite is true. Our fixation on test results deflects attention from fundamental educational problems and so hinders reform (National Commission on Testing and Public Policy, 1990).

In this paper we examine four aspects of current high-stakes testing that impact on minority students and others traditionally poorly served by the education system based primarily on research done at Boston College over the past 30 years. After reviewing the available evidence, we come to the following conclusions:

1. High-stakes, high-standards tests do not have a markedly positive effect on teaching and learning in the classroom
2. High-stakes tests do not motivate the unmotivated
3. Contrary to popular belief, "authentic" forms of high-stakes assessments are not a more equitable way to assess the progress of students who differ in race, culture, native language, or gender[2]
4. High stakes testing programs have been shown to increase high school dropout rates – particularly among minority student populations

Before developing these points, it is worth looking at some of the trend data available on the test performance of minority students over the last 30 or so years. Results from several studies including the National Assessment of Educational Progress (NAEP), the National Educational Longitudinal Survey (NELS), and the Scholastic Assessment Test (SAT) indicate that while the gap between black and Hispanic students and their white counterparts has been generally narrowing over the years it is still, as Campbell and his colleagues (1997) put it, "a substantial" one (p. 67).

Some illustrations from the mathematics components of these tests will suffice. For example, results from the 1996 trend NAEP provide

evidence that averages achieved by 9- 13- and 17- year-old Black and Hispanic students in mathematics were substantially higher than those achieved by their counterparts in 1973 (Figures 1-3). In addition, gains made by these groups were generally larger than those made by White students (Campbell et al. 1997). However, tempering this positive trend is the realization that the average proficiency for White 13-year-olds as measured on the 1996 NAEP scale was around the level achieved by Black 17-year-olds.[3] In addition, data from the main NAEP for 1996 indicates that at grades four, eight, and twelve achievement gaps between White students and their Black and Hispanic counterparts ranged from .8 to 1.1 standard deviation units. To put this in context, a difference of one standard deviation implies that just 16% of the low achieving group perform at levels exceeded by 50% of the higher achieving group. There is also evidence that the achievement of Native Americans in mathematics on the main NAEP was substantially below that of Whites at all grade levels and that the difference increases as students move through the school system.

<Insert Figures 1, 2 and 3 about here>

Similar outcomes for racial differences in mathematics are found on longitudinal studies such as NELS and High School and Beyond (HSB) (Figure 4). Outcomes by race on the college admissions tests also confirm these findings. As the trend lines in Figures 5 and 6 indicate, Asian American students were the highest performing group on both the American College Testing (ACT) assessment and the SAT over the 20 years spanning 1977-1997. White students, the second highest performing group, achieved at average levels that were about .3 of a standard deviation unit lower than Asian students. The average performance of Black students was consistently lower than all other groups and stood well below that of Asian and White students in 1997. Nevertheless, it is important to emphasize that while all groups improved their performance from the mid 1970s on, some of the largest improvements on average were associated with Blacks. The improved

performance of Mexican American students, especially on the ACT assessment, is also noticeable. An important finding of the McLure, Valiga and Sun (1997) study of ACT scores between 1987 and 1996 was that the improved performance of minority groups in general on the ACT mathematics test coincided with a pattern of increased course taking by students from these groups.

<Insert Figures 4, 5 and 6 about here>

This educational achievement gap is hardly news. It is a well-established fact that, using almost any measure, black students nationwide do not perform as well as whites. However, less well established are the reasons for this fact. The nature of inquiries into the achievement gap for racial/ethnic groups have run the full gamut. A lot of attention has been placed on the test itself and possible flaws therein. However in today's high stakes testing context, it is becoming harder and harder to blame the test for large performance differences between groups -- what the courts call adverse impact. Now there is a realization that differential results can reflect genuine group differences in whatever trait is being assessed (Linn & Bond, 1994; Jencks & Phillips, 1998). These tests – even when faulty -- are signaling something is wrong. Lani Guinier's analogy of test results being the miner's canary is apt (Guinier& Sturm, 1996)[4]. Further, there will be no breakthrough in the technology of testing to ameliorate the adverse impact we see on cognitive measures. Technically we can only tinker at the margins.

There is one caveat about the malleability of the testing technology. The Achilles' heal of the high-stakes high-standards assessment movement is that the standard setting process increasingly results in standards -- cut scores -- that are very high relative to current distributions of scores on standardized achievement tests. Many policy makers fail to recognize that the validity issue cannot be separated from the choice of the cut-score, which automatically triggers a decision or inference. A classic example of this disregard is the four cut-scores used in NAEP to classify mathematics attainment. The NAEP percentages of

students at the advanced level never exceed 4%. However, this NAEP description of national math attainment is a classic ipse dixit. The bleak picture is valid only so long as one accepts the NAEP cut scores as definitive. Independent achievement data from the SAT, ACT and Advanced Placement tests in mathematics call into question the NAEP percentages.[5] Another example is the Massachusetts Comprehensive Assessment System (MCAS) which results in large numbers of students falling into the "failure" and "needs improvement" category despite contradictory standardized norm referenced data. In short the admittedly arbitrary nature of cut-score methodologies cries out for the external validation of the categories produced and used to label students.

There are three additional problems with very high cut scores (Koretz and Barron, 1998). First, they result in schools and parents getting essentially no information other than, "failure" "inadequate" or "needs improvement" about large numbers of children. Second they provide a misleading view of change. Large changes that stay within one of the levels go unnoticed, and trivial changes that cross a cut-score boundary are treated as important. Finally, the school systems often set goals for improvement-- either total improvement or an annual rate -- that are simply not possible to meet by legitimate means. This gives teachers an additional incentive to teach to the test in inappropriate ways.

The Impact of High-Stakes Testing on Teaching and Learning

Contemporary policymakers who advocate the use of tests as levers of educational reform certainly recognize the historical role of testing in controlling what is taught and learned. There seems to be little argument that tests affect the curriculum; it is almost an educational truism. Indeed, the power of an examination to shape what is taught and learned was noted at least as far back as the 16th century, when Philip Melancthon, a Protestant German teacher, wrote in *De Studiis Adolescentum*, "no academical exercise can be more useful than that of

examination. It whets the desire for learning, it enhances the solicitude of study while it animates the attention to whatever is taught" (Hamilton, 1853, p. 769, cited in Madaus & Kellaghan, 1992, p. 121).

The concept of the power of an important test is beautifully captured by chief inspector of schools Edmond Holmes writing about 19th century school examinations in Great Britain. Victorian style apart, Holmes' observation remains true today and for the United States:

> Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion, as we tend to value the results of education for their measurableness, so we tend to undervalue and at last ignore those results which are too intrinsically valuable to be measured. (Holmes, 1911, p. 128.)

How do important tests exert such an influence? What are the mechanisms that vest high stakes tests with their power to change instruction and learning? There are four principles that explain the importance of such tests on what is taught and learned.[6]

1. The more any quantitative social indicator is used for social decision-making, the more likely it will be to distort and corrupt the social process it is intended to monitor. This very general principle comes from the work of Don Campbell (1975) and is a social version of Heisenberg's uncertainty principle.[7] The following three principles are special cases of this principle playing out in education.

2. If teachers perceive that important decisions are related to the test results they will teach to the test. One of the necessary conditions for measurement-driven instruction to work is that valued rewards or serious sanctions are perceived to be triggered by test performance.[8]

3. When test stakes are high, the tradition of past exams comes to define the curriculum. Once a high-stakes testing program has been in place for several years teachers see the kind of intellectual activity required by the previous test questions and prepare students to meet these demands.[9]

4. When teaching to the test, teachers pay attention to the form of the test as well as the content. When teaching to the test the form of the questions can narrow the focus of instruction, study, and learning to the detriment of other skills. [10]

Today advocates of "authentic" assessment presume that such assessments are somehow outside the purview of the three principles. Not so. A powerful illustration that supply-type exams – exams that require students to produce their own answer to a question rather than choose from a set of possible answers -- are also likely to distort instruction comes from samples of student essays from an Irish examination from the mid-1940s, the Primary Leaving Certificate (Madaus & Greaney, 1985: Madaus, 1988). The example shown in Figure 7 illustrates how students learned to memorize stock responses that would be adaptable to any writing prompt. Note the similarity between the answers to three different writing prompts from three different students across three different years. Scores from such tests said more about students' memories and test-taking strategies than they did about students' ability to write.[11]

[Insert Figure 7 about here]

How do these four principles affect minority students? A 1992 NSF sponsored national study of the effects of high-stakes tests – by and large standardized multiple choice tests -- on math and science instruction found that students in high-minority classrooms are affected significantly more by such tests than are their peers in low-minority classrooms.[12] Teachers with more than 60% minority students in their class, compared to teachers with less than 10% minority students,

reported more reliance on mandated standardized tests for various uses, more test pressure, and more test preparation and influence on instruction. They more often reported that test scores were "very" or "extremely important" to either themselves or administrators for placement in special services, determining graduation, recommending textbook, planning curriculum and instruction, evaluating student progress, and giving feedback to students.

About 75% of both math and science teachers with high-minority classes reported pressure from their district to improve standardized test scores, in comparison with about 60% of teachers with low-minority classes. Teachers in high-minority classrooms significantly more often reported teaching test-taking skills, teaching topics known to be on the test, increasing emphasis on tested topics, beginning preparation more than a month before the test, and including topics not otherwise taught. Finally the case studies conducted in six urban districts with large minority enrollments confirmed the finding of the national survey that instruction by teachers facing high stakes testing pressure is often heavily oriented toward test preparation. There is little question that high-stakes tests have greater consequences for minority and poor children than they do for majority and more affluent students, albeit they too are impacted by the power of these tests.

The Impact of High-stakes Testing on Student Motivation

The idea that mandating a high-stakes assessment -- national or state level -- would improve student motivation to learn, is one that goes back at least to the 18th century.[13] A recent American Educational Research Association (AERA)-sponsored monograph examined the claim that high-stakes assessments will move the obstinate, dispirited, lazy, or recalcitrant students to try harder in school (Kellaghan, Madaus and Raczek, 1996). Three corollaries associated with contemporary assertions about the motivational force of high-stakes emerged.[14] First, the greater the reward offered (or the more noxious the consequences of not

complying), the harder students will try. Second, the meaning of rewards and punishments is essentially the same for all students (poor, middle-class, and minority). And third, student arousal is maximized when rewards are distributed on a competitive basis.[15]

The concept of motivation used in claims about the power of examinations lacks definition and clarity (Kellaghan et al. 1996). The complexity of the construct, its variety of meanings, and the different mechanisms that elicit it, are not adverted to in blanket claims about the examinations' motivational power. Investigators in cognitive psychology, while identifying goals as central in the motivational process, at the same time point to the complexity of the relationships and processes involved, the idiosyncrasy of individuals' choice of goals and subgoals, and differences between students in their assessment of their ability and self-efficacy.[16] There seems to be little appreciation among reformers that the construct may be even more complex in the context of external high-stakes national examinations embedded as they will be in complex school, cultural, and social networks. In other words, advocates of the motivational potential of examinations have not paid enough attention to who will be motivated and who will not, a point that is particularly relevant when the examinations are referenced to "world class" standards that all students, regardless of grade level, circumstances, context, and individual differences, are expected to attain.

Analysis of the motivation process indicates that students must first see that striving for rewards attached to examination performance is not only important in their lives but is realistic (Kellaghan et al. 1996). In practice, however, some students immediately dismiss the examination because they perceive the rewards to be unobtainable. Others feel they lack the ability to do what is necessary to pass. Again, others, while believing that they may have the ability to pass, are not motivated to work toward examination success because they do not see the test credential as necessarily resulting in jobs or college, because of scarcity, competition, or lack of relevance in their social setting. Still others

perceive not only that they lack the ability, but that the rewards in reality are illusory even for many who might pass. Further, there is no reason to believe that the motivational power of examinations will be the same at all grade levels. Since present reform proposals call for testing at elementary and middle grades, we need to ask how students at these age levels will perceive examinations which are distant in time in terms of directing their behavior.

There can be no doubt that some students in other countries do indeed work hard to pass high-stakes external examinations. Further, many pay not inconsiderable sums of money to attend test-preparation schools and some, no doubt, internalize the competitive values embedded in the examinations. It may even be that some students develop genuine intrinsically motivated behavior or at any rate relatively autonomous forms of extrinsic motivation. However, experience with external examinations would seem to indicate that all too many students focus their efforts on mastering strategies to help them over the examination hurdle rather than on developing mastery of subject matter and honing lasting competencies. This is a consequence of external examinations that does not seem to have been anticipated by some of its proponents in the United States. For example, a laudable objective of the Learning Research and Development Center and the National Center on Education and the Economy (n.d.) is that their proposed examination system would lead students to see that "school is a place to learn and become competent, not just to be labeled as smarter or slower than others". While implicitly recognizing the distinction between competence and performance, there does not seem to be an appreciation that high-stakes external examinations are likely to thwart rather than support the attainment of this objective. Wilfred Sheed's (1982) description of a cramming school offers an amusing insight into the effects of focusing on performance rather than on learning in the context of external examinations. The crammer, Jenkins Tutorial Establishment in London,

> offered... successful examination results as it might a forged passport.. [bypassing] education altogether. Their only texts were examination papers-- all the relevant ones set in the last fifty years, with odds of repetition calculated and noted as in the Racing Form. Within six months, I was able to pass London matriculation without knowing any of the subjects involved; and by applying Jenkins' method later, to pass every exam that ever came my way afterwards. Hence I remain a profoundly uneducated man (p.117).

There are further consequences of external examinations (Kellaghan et. al, 1996). For example, many teachers, under pressure to help students secure good examination results, will be more controlling in their teaching. In fact, the "rhetoric from Washington continues to advocate greater accountability, greater discipline, and increased use of standardized testing, all of which are means of exerting greater pressure and control on the educational process" (Deci, Vallerand, Peletier and Ryan, 1991, p.342). The fact that examinations are top-down and bureaucratically controlled, rather than bottom-up and under the control of practitioners, parents, and the local community, has implications for goal adoption and pursuance. Further, insofar as it involves increased control of teacher and student behavior through the imposition of rewards and sanctions associated with test performance, there would appear to be the possibility of other undesirable consequences. When controlling events are perceived to determine behavior, the student's needs for competence, self-determination, conceptual learning, and creativity will not be met but rather diminished.[17]

Examinations can also lead to an increase in competition and cheating on the part of some students. We argued above that situations that put emphasis on test performance rather than on learning per se narrow the curriculum to what is embodied in the tradition of past examinations. We therefore expect the pursuit of high stakes examination performance in time to corrupt the examination with the

result that inferences about achievement from performance will no longer be valid.

A further important consideration is that students, who may not be motivated to pursue examination success for socio-cultural reasons, or on the basis of estimates of their present levels of achievement or ability, are likely to become alienated, not only from examinations but from the whole educational process (Ogbu, 1991; Steele, 1997). The fact that around ten percent of students, mostly from disadvantaged backgrounds, avoid taking any public examination before leaving school is regarded as a serious problem in several European countries. In America, research on dropouts also shows that many, while fully appreciating the importance of educational credentials, do not believe that such credentials are of much help in their social milieu. The motivational argument of proponents of high-stakes national examinations does not address these realities. We will examine the drop out issue below. Suffice it to say here that motivation and dropout issues are intertwined.

It would seem that the use of external motivation techniques in industry has little relevance for the high-stakes assessment situation in education. In industry, skills have already been acquired, rewards are tangible and immediate and serve to reinforce and direct behavior. Further, the feedback mechanisms in industry are well developed and more immediate. None of these conditions hold for high-standards high-stakes assessments. Another high-stakes exam situation -- no-pass no-play or no-pass no-drive programs -- seem not to work very well despite the fact that they are more immediate and real for many students than are high-stakes school assessments (Kellaghan et al. 1996).

These considerations lead to the conclusion that it is incumbent on reformers to weigh more carefully than they have done up to this point the costs and benefits that are likely to be associated with the variety of outcomes of examination-induced motivation (Kellaghan et al 1996). While there are possible positive aspects to the examination movement,

for example in the specification of clear goals and standards for education for teachers and students, greater consideration needs to be given to the nature of those goals and standards and how they might need to vary for different students. Other important issues will also have to be addressed. These include the probability that many students will not be motivated at all. Moreover, for many who are motivated, the high-stakes associated with the assessment may work toward focusing their efforts on improving their test performance rather than on the more demanding job of developing general competence, higher-order thinking skills, improved problem-solving ability, and creativity. It is these latter traits which reformers claim will be the eventual outcome of standards based reform. Until such issues are addressed—and even if they were -- we can have little confidence in high-stakes assessments as a panacea for the ills for American education.

The Impact of Authentic Assessments on the Performance of Students who Differ in Race, Culture, Native Language, or Gender

An outgrowth of the high-stakes, high-standards movement is the belief that "authentic" assessments are more equitable for assessing the progress of students who differ in race, culture, native language, or gender (Wiggins, 1989).[18] In 1992 the United Kingdom (UK) provided a type of naturally occurring experiment, albeit in another country, that allowed the examination, at least from the perspective of initial implementation, of a number of such equity or fairness claims made for "authentic" assessments with high-stakes attached.[19]

The UK data permitted an examination of the relative performance of gender, linguistic, low income, and special needs groups on the high-stakes "authentic" assessments administered in 1992 to seven year olds as part of the national curriculum (Thomas, Madaus and Raczek, 1998).[20] It was found that irrespective of the method of "authentic" assessment and once all other factors had been taken into account, there were substantial differences between particular groups of students defined in terms of gender, low income, special needs and native

language. The subject level results for English, mathematics and science attainment showed that students from low income families, or whose first language was not English, or who were special needs students performed at a significantly lower level than all other students. Gender was the only factor that varied in impact across the three subjects. In English and mathematics girls performed at a significantly higher level than boys, although the difference in mathematics was small. However in science there was no significant difference between boys and girls.

As would be expected, student age had a positive impact with older students performing better than younger students. Regarding the higher attainment of older students an episode in the 1980s occasioned by high-stakes testing in the United States is worth mentioning. One tactic used by schools to improve high-stakes reading scores in upper primary grades was to put pressure on primary teachers to begin to emphasize reading skills more strongly. Such an emphasis however, came at the expense of other more traditional goals for children in K-3. A related tactic to improve upper level performance was that of "red shirting" kindergarten students. Educators realized that retaining children in kindergarten or not letting them enter in the first place because they do not have the necessary "readiness" should, in the long run, contribute to higher test scores if for no other reason than students are a year older when they take the high-stakes tests.[21] The English data validate the fact recognized in the 80s that improved test performance is related to age.

It is important to keep in mind that the English results are from the early stages of a new program, from a different country and educational system, and from seven-year-olds only. Nonetheless, at the very least, they point to the need to carefully monitor group differences in light of positive equity claims made about the "authentic" assessment technology. These results are also cautionary to those making such equity claims.

The Impact of High-Stakes Assessments on High-School Dropout Rates

One aspect of high-standards high-stakes assessments that needs considerably more attention than it has received to date is the relationship between such assessments and dropping out of high school. There are five suggestive lines of evidence that argue for a careful examination of this relationship.

The first intriguing data comes from the Minimum Competency Testing (MCT) era. An overall, albeit quite crude, view of the relationship between MCT and dropout rates comes from examining the ten states with the highest 1986 dropout rates and the ten states with the lowest dropout rates (Kreitzer, Madaus and Haney, 1989). Correlational data indicate a strong relationship between attrition or dropout rates and the existence of MCT programs in these states. [22] Half of the ten states with the lowest 1986 dropout rates had no minimum competency testing programs. The other five states with low dropout rates had MCT programs that could be characterized as involving relatively low stakes: four used the tests for decisions about remediation; one used them only for accountability. None required the tests for critical decisions about graduation or grade promotion. Furthermore, in three of these five states, local, not state, education agencies set the standards.

States with the highest dropout rates, on the other hand, had MCT programs where standards were set, at least in part, at the state level. Nine of the ten used the tests in decisions about high school graduation; four used them in decisions about promotion. In sum, these ten states with the highest dropout rates employed minimum competency tests with higher stakes and less flexible standards than the states with the lowest dropout rate.

These data are not evidence of a causal relationship between high-stakes MCT programs and dropout rates. The states with the highest dropout rates differed in obvious ways from the states with the

lowest dropout rates. The latter were largely western and midwestern, and they had a relatively low representation of minority and poor students among their school-age populations. Perhaps high dropout rates are symptoms of the educational system's failure that spurred legislators to mandate MCT programs in the first place or perhaps MCT does contribute in some way to the dropout problem (Kreitzer et al, 1989). In any case, crude as the Kreitzer et al data may have been, they underline the need to explore further competing hypotheses about the connection or lack thereof between high-stakes testing and dropping out.

A second bit of implicative, albeit correlational data on dropouts and MCT comes from an examination of the relationship between MCT in eighth grade and early high school dropout patterns (Reardon, 1996).[23] The results suggest that in schools with high proportions of low-SES students, MCTs are linked to much higher dropout rates. The dropout rates from these schools are 2 to 6 percentage points higher, on average, than from similar schools with no such requirement. The overall conclusion is that "it is the concentrated poverty of these schools and their communities, and their concomitant lack of resources, that link MCT policies to higher dropout rates, rather than other risk factors, such as student grades, age, attendance, and minority group membership." (p.5)

A third, more recent line of suggestive data on the possible connection between dropping out and high-stakes assessments, particularly as it relates to minorities, comes from an analysis of data from the Texas Assessment of Academic Skills (TAAS) (Fassold, 1996).[24] This study found that due to the TAAS requirement over 25,344 African American and Hispanic students of Texas's 1993 sophomore cohort dropped out of school. For white students the drop out figure was 14,809. African American and Hispanic students dropped out at a significantly higher rate than did white students even controlling for SES[25], academic track, language program participation, and school quality. While this data applies to one particular cohort, it fits well with

patterns our colleague Walt Haney has observed in enrollment rates in Texas over a 20-year period. After breaking down enrollment rates by ethnicity, Haney found that the ratio of high school graduates to ninth graders three years earlier has dipped considerably for both black and Hispanic students since about 1990 but has remained quite stable for white students. For example, in 1989 (the year prior to TAAS), the ratio of high school graduates to ninth graders three years earlier for whites was around .76. Even after TAAS was implemented in the 1990-1991 school year, this ratio never dipped below .73. In 1989, the ratio of high school graduates to ninth graders three years earlier for blacks was quite close to that for whites - at around .74. However, since the implementation of TAAS, this ratio has dipped dramatically - falling to a low of around .55 in 1994 and hovering around .57 in 1997. Hispanic students also experienced a similar drop in their ratio during this time period. Again, we cannot argue causation from these correlational data. These findings call out for more detailed interviews with a carefully chosen random selection of drop outs to gauge the actual impact of the TAAS on their decision to leave school.

A fourth line of evidence comes from work on the relationship between grade retention, being overage-for-grade and drop-out rates. Research on the effects of grade retention – whether or not it is coupled with high stakes testing - has generally concluded that its harms outweigh any purported benefits (Darling-Hammond and Falk, 1997; Smith and Shepard, 1989). In particular, there are negative effects associated with a student being overage-for-grade as a result of having been retained in an earlier grade. Research has found that being overage for grade eats away at students' sense of efficacy, with the impact especially severe for African-American students. Compared to on-grade students, these overage students are twice as likely to be retained in grade again (Texas Education Agency, 1996). Many of these students ultimately become disengaged and drop out; and many, under pressure of high-stakes testing, drop out earlier in their school careers (Wehlage

and Rutter, 1986). In fact, being overage for grade is a better predictor of dropping out than below-average test scores (Texas Education Agency, 1996). These findings were played out fifty years ago in Ireland in response to a mandatory primary-school leaving certificate examination that was administered to all sixth-grade pupils in Ireland between 1943 and 1967 (Madaus and Greaney, 1985). Teachers were employing a policy of not promoting weaker pupils in order to control the potential failure rate on the examination. This non-promotion policy tended to take place at two points in the system – grades 3 to 4, and grades 5 to 6. A pupil that was held back at one or both points would have been eligible to leave school before they reached the sixth grade – lessening the number of these overage students that ever sat for the primary examination.

A final piece of related – but little known -- evidence about the national picture regarding dropouts comes from the recent Third International Mathematics and Science Study (TIMSS) data which indicates that about 35% of the senior cohort was no longer enrolled in school when the TIMSS tests were administered in the Spring of the senior year (Mullis et al, 1998).[26] Further, a post hoc check revealed that the NAEP data on dropouts are quite consistent with the TIMSS data (see Appendix A for details). The no longer enrolled figure for TIMSS is quite a bit higher than the NCES dropout rate of 12% for the United States for 1995. One explanation for this discrepancy is that many students who are enrolled in October, the month used in calculating the NCES rate, are no longer in school by April/May when the TIMSS tests were administered. This TIMSS/NAEP higher rate of missing seniors comes at a time when all but one state has a state testing program, and when increasingly the state tests are tied to graduation. The TAAS results discussed above are a case in point. Once again this conjecture is correlational not causal but it points to the need for a more carefully designed study to gauge the impact of high-standards high-stakes assessments on decisions to drop out of school. For example, dropout

statistics obviously are time sensitive. It would be helpful to track attrition between October and April to sort out when, but also why, students leave in their senior year.

Conclusion

Testing is a powerful, but often blunt, tool. Like a medication, it may fail or have diverse, unintended negative consequences. A testing program, for example, may unfairly or unreasonably deny opportunities to classes of people; it may reward or punish the wrong individuals or institutions; and it may undermine the performance of institutions it is intended to strengthen. Further, the four issues described above, (1) the impact of testing on teaching and learning; (2), the motivational power of high-stakes assessments; (3) claims that "authentic assessments" are fairer to minority students than traditional standardized multiple choice testing; and, (4) the role of high-stakes assessments in decisions to drop out of school all need to be carefully monitored. Policymakers and test users have been able to turn to extensive commercial, not for profit, and governmental infrastructures that have evolved over the past 90 years to assist them in test development, administration, scoring, and reporting. However, there has been no analogous infrastructure for independently evaluating a testing program before or after implementation, or for monitoring test use and impact.

In its 1990 report *From Gatekeeper to Gateway: Transforming Testing in America*, the National Commission on Testing and Public Policy (NCTPP) recognized "the need for sound, fair, and reasonably efficient mechanisms to help make difficult decisions about individuals and institutions."(p. 2). It also noted that "although tests have become important instruments of public accountability, there are few mechanisms to *audit* or *appraise* the quality of publicly sponsored tests, to monitor their use as instruments of social policy, and to assess their impact on individuals, groups, and institutions." (p.31, emphasis in original). To remedy this the Commission called for "the development of *additional institutional means* to examine the quality of tests and

assessment instruments and to provide oversight of test use."(p.13, emphasis added)[27]

Since 1990, a variety of factors, including changes in statute and case law, national and state policy initiatives, and plebiscites, have both transformed educational testing and increased the need for oversight. All have altered the uses of tests, in some cases fundamentally, and all have the potential to affect profoundly both educational systems and individuals, particularly students currently ill-served by the educational system -- economically disadvantaged students, students with disabilities, students for whom English is a second language, and students from ethnic minorities. These changes also raise difficult equity and technical questions, placing policymakers, educators, and even researchers in uncharted territory. Perhaps at no time in the past half-century have questions and consequences surrounding educational testing been as widespread and serious.

In many other areas where technology and policy intersect, the public insists on oversight — including technical oversight — to protect individuals from unintended negative effects. For example, faced with the policy decision to introduce a major new untried medical technology to millions of children, particularly a treatment that would be given to healthy children as well those who were ill, the public would ask about the safety, efficacy, quality, and social and economic effects of the new technology or treatment, and public agencies have been established to address such concerns systematically. The effects of testing are now so diverse, widespread, and serious that it is necessary to establish mechanisms for catalyzing inquiry about, and systematic independent scrutiny of them.

In response to the need for monitoring high-stakes educational testing, in September of 1998 the Ford Foundation provided start-up funding for an independent, institutional oversight agency called the National Board on Educational Testing and Public Policy (NBETPP). The NBETPP will review testing programs and catalyze close consideration of

the diverse uses of testing in education. The Board will be a permanent institutional entity -- not a transitory commission that studies a situation and then issues a report and recommendations, which along with the commission itself is easily ignored and then soon forgotten (Bell, 1997).

Few would disagree with the objectives of raising educational standards and in general improving the quality of American education. It seems fairly clear, however, that efforts to foster academic achievement should involve more than simply setting demanding standards and mandating examinations that are referenced to them. The task remains of identifying strategies to achieve the desirable reform objectives, efficiently and effectively, and without having a negative impact on any subpopulation of students. Those strategies will, among other things, need to address the issue of restructuring the academic experiences of students in ways that will help them appreciate the value of academic achievement, increase their expectations and aspirations, and enhance their sense of academic efficacy.[28] This is a much more difficult and, we dare say, expensive task than mandating tougher standards and an external examination referenced to them.

References

Berliner, D. C. (1993). Mythology and the American system of education. *Phi Delta Kappan*, 75(8):632-640.

Biemiller, A.J. (1996). Aid to elementary and secondary education. *Hearings before the General Subcommittee on Education of the Committee on Education and Labor. House of Representatives, 89th Congress, 1st Session, on H.R. 2361 and H.R. 2362, Jan. 29, 1966, Vol. I.*

Bracey, G. W. (1992). The second Bracey report on the condition of public education. *Phi Delta Kappan*, 74(2):104-117.

Bracey, G. W. (1993). The third Bracey report on the condition of public education. *Phi Delta Kappan*, 75(2):104-117.

Brooke, N. & Oxenham, J. (1984). The influence of certification and selection on teaching and learning. In J. Oxenham (ed.), *Education versus qualifications* (pp. 147-175). London: Allen & Unwin.

Cannell, J. J. (1987). *National normed elementary achievement testing in America's public schools: How all fifty states are above the national average.* Albuquerque, NM: Friends for Education.

Cannell, J. J. (1989). *The "Lake Wobegon" report: How public educators cheat on standardized achievement tests.* Albuquerque, NM: Friends for Education.

Campbell, D. T. (1975). On the Conflicts Between Biological and Social Evolution and Between Psychology and Moral Tradition. In *American Psychologist* 30(12): 1103-1126.

Campbell, J. R., Voelkl, K. E., and Donahue, P. L. (1997). *NAEP 1996 trends in academic progress.* Washington, DC: National Center for Education Statistics.

Celebrezze, A. (1965). *Hearings before the Committee on Education and Labor, House of Representatives, Education Act of 1965, H.R. 2361 and H.R. 2362.* 89th Congress. Washington, D.C.: U.S. Government Printing Office.

Darling-Hammond, L., and Falk, B. (1997). Using Standards and Assessments To Support Student Learning. *Phi Delta Kappan*, 79(3): 190-99.

Deci, E. L. & Ryan, R. N. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum.

Deci, E. L., Spiegel, N. H., Ryan, R. M., Koestner, R., & Kauffman, M. (1982). Effects of performance standards on teaching styles: Behavior of controlling teachers. *Journal of Educational Psychology* 74, 852-859.

Deci, E. L., Vallerand, R. J., Peletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist*, 26, 325-246.

Dore, R. P. (1976). *The diploma disease: Education, qualification, and development.* London: Allen & Unwin.

Dweck, C., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256-273.

Fassold, M. A. (1996). *Adverse racial impact of the Texas Assessment of Academic Skills.* San Antonio TX: Mexican American Legal Defense and Education Fund.

Guinier& Sturm, (1996). The future of affirmative action: Reclaiming the innovative ideal. California Law Review, 84(4), 953-1036.

Hamilton, W. (1853). *Discussions in philosophy and literature, education, and univerisity reform,* 2nd edition. London: Longman.

Hamilton, L.S., Nussbaum, E.M., and Snow, R.E., "Interview Procedures for Validating Science Assessments," *Applied Measurement in Education*, vol. 10(2), 1997, pp. 181-200.

Holmes, E. G. A., (1911). *What is and what might be: A study of education in general and elementary in particular.* London: Constable.

Jencks, C. (1998). Racial bias in testing. In C. Jencks and M. Phillips (eds). *The black-white test score gap.* Washington, DC: Brookings Institution Press.

Jencks, C. & Phillips, M. (eds.) (1998). *The black-white test score gap.* Washington, DC: Brookings Institution Press.

Kellaghan, T., Madaus, G. F., & Raczek, A. E. (1996). *The use of external examinations to improve student motivation.* Washington, DC: American Educational Research Association. Monograph.

Koretz, D., and Barron, S. I. (1998, in press). *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica: RAND.

Kreitzer, A. E., Madaus, G. F., and Haney, W. M. (1989). Competency Testing and Dropouts. In L. Weis, E. Farrar & H. G. Petrie (eds) *Dropouts from School: Issues, Dilemmas, and Solutions*, pp. 129-152. Albany, NY: State University of New York Press.

Linn, R. & Bond, L. (1994). *Studies of the extent of adverse impact of certification rates on federally protected groups and the extent to which assessment exercises are free of bias and unfairness: A report submitted to the National Board for Professional Teaching Standards.* Detroit, MI: National Board for Professional Teaching Standards.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. Tanner (ed.), *Critical Issues in Curriculum*, 8-121. Chicago: University of Chicago Press.

Madaus, G. F. (1993). A national testing system: Manna from above: An historical/technological perspective. *Educational Assessment, 1*(1), 9-26.

Madaus, G. F. (1995). *Do we have a crisis in education? The fashioning and amending of public knowledge and discourse about public schools.* Division D American Educational Research Association, Vice-presidential address, presented at the annual meeting, April 17, 1995, San Francisco CA.

Madaus, G. F. & Greaney, V. (1985). The Irish experience in competency testing: Implications for American education. *American Journal of Education*, Vol. 93, 268-294

Madaus, G. F. & Kellaghan, T. (1991a). *Student examination systems in the European community: Lessons for the United States.* Contractor report submitted to the Office of Technology Assessment, United States Congress.

Madaus, G. F., and Kellaghan, T. (1991b). National testing: Lessons for America from Europe. *Educational Leadership* 49(3), 87-93.

Madaus, G. F., & Kellaghan, T., (1992). Curriculum evaluation and assessment, p.121. In P. W. Jackson (ed.), *Handbook of research on curriculum*, 119-54. New York: Macmillan.

Madaus, G. F., West, M. M., Harmon, M., Lomax and Viator, K. A. (1992) and West, M. M., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12: Appendix D: Testing and teaching in six urban sites.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy; Boston College.

Marland, S. (1965). Hearings before the Committee on Education and Labor, House of Representatives, Education Act of 1965, H.R. 2361 and H.R. 2362. 89th Congress. Washington, D.C.: U.S. Government Printing Office.

McLure, G. T., Sun, A., and Valiga, M. J. (1997). *Trends in advanced mathematics and science course-taking and achievement among ACT-tested high school students: 1987-1996.* Iowa City: American College Testing Program.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York: The Free Press.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics and science achievement in the final year of secondary school: IEA's third international mathematics and science study.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.

National Commission on Testing and Public Policy, (1990). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: National Commission on Testing and Public Policy.

National Institute of Education. (1981, June). *Minimum competency testing clarification hearings.* Hearings held June 8, 9, 10 in Washington D.C. (ERIC Document Reproduction Service, Documents ED215000, ED215001, ED215002)

Ogbu, J. (1991). Immigrant and involuntary minorities in comparative perspective. In J. Ogbu and M. Gibson (eds). *Minority Status and Schooling.* New York: Garland.

Popham, W. J. (1983). Measurements as an instructional catalyst. *New Directions for Testing and Measurement* 17:19-30.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, 68, 680-682.

Popham, W. J., Cruse K. L., Rankin, S. C., Sandifer, P. D.& Williams, P. L. (1985). Measurement-driven instruction: It's on the road *Phi Delta Kappan*, 66, 628-635.

Reardon, Sean F. (April 8, 1996). *Eighth Grade Minimum Competency Testing and Early High School Dropout Patterns.* Paper presented at the Annual Meeting of the American Educational Research Association. New York.

Shepard, L. A., & Smith, M. L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. *Elementary School Journal* 89:135-145.

Shohamy, E. (1993). *The power of test: The impact of language tests on teaching and learning* Occasional Papers. The Natal Foreign Language Center, John Hopkins University, Washington DC.

Sheed, W. (1982). *Transatlantic blues.* New York: Dutton.

Smith, A. (1990). *An inquiry into the nature and causes of the wealth of nations.* (2nd ed.). Chicago: Encyclopaedia Britannica.

Shepard, L. A., and Smith, M. L. (1989). Academic and emotional effects of kindergarten retention in one school district. In L. A. Shepard and M. L. Smith (eds.) *Flunking Grades: Research and Policies on Retention*, pp. 79-107. London, England: Falmer Press.

Spaulding, F. T. (1938). *High school and life: The Regent's inquiry into the character and cost of public education.* New York: McGraw Hill.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist* 52(6):613-629.

Tanner, D. (1993). A nation 'truly' at risk. *Phi Delta Kappan*, 75(4):288-297.

Texas Education Agency. (1996). *Comprehensive Bienniel Report on Texas Public Schools: A Report to the 75th Texas Legislature.* Austin: Texas Education Agency.

Thomas, S., Madaus, G. F., & Razcek, A. (1998). Comparing teacher assessments and standard task results in England: The relationship between pupil characteristics and attainment. *Assessment in Eduaction: Principles, Policy & Practice*,5(2), in press.

Webb, F. R., Covington, M. V., and Guthrie, J. W. (1993). Carrots and sticks: Can school policy influence student motivation? In T. M. Tomlinson, (ed.) *Motivating students to learn: Overcoming barriers to high achievement*, pp. 99-124. Berkeley, CA: McCutchan.

Wehlage, G. G., and Rutter, R. A. (1986). Dropping Out: How Much Do Schools Contribute to the Problem? *Teachers College Record*: v87 n3 p374-92.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 70(9), 703-713.

Figure 1. Trends in Racial/Ethnic Group Differences: NAEP Mathematics - 9-year-olds.

Note: The performance of White students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.

Sources: "Weighted Means, Standard Deviations, and Percentiles of Mathematics Distributions with Jackknifed Standard Errors" (Princeton, NJ: Educational Testing Service, undated and unpublished tabulations); John Mazzeo, Educational Testing Service, personal communication, (September 1997); Campbell, Voelkl and Donahue (1997).

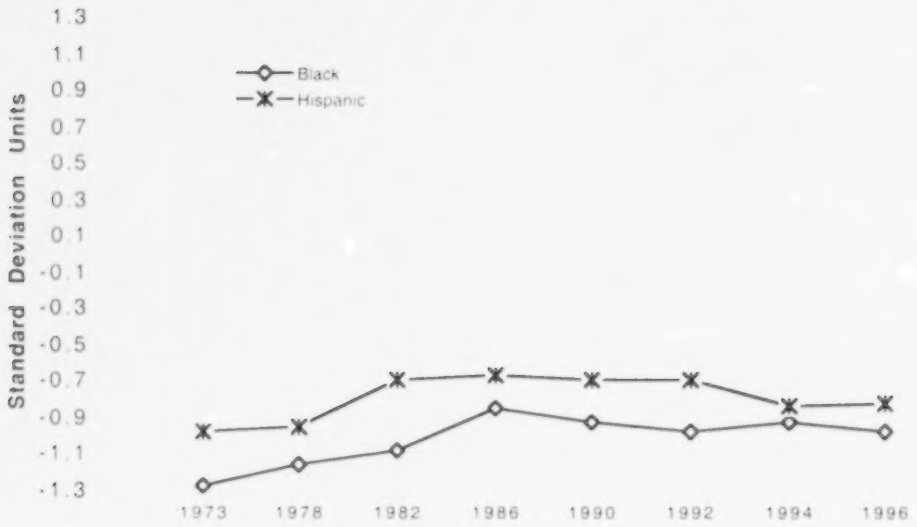

Figure 2. Trends in Racial/Ethnic Group Differences: NAEP Mathematics -13-year-olds.

Note: The performance of White students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.

Sources: "Weighted Means, Standard Deviations, and Percentiles of Mathematics Distributions with Jackknifed Standard Errors" (Princeton, NJ: Educational Testing Service, undated and unpublished tabulations); John Mazzeo, Educational Testing Service, personal communication (September 1997); Campbell, Voelkl and Donahue (1997).

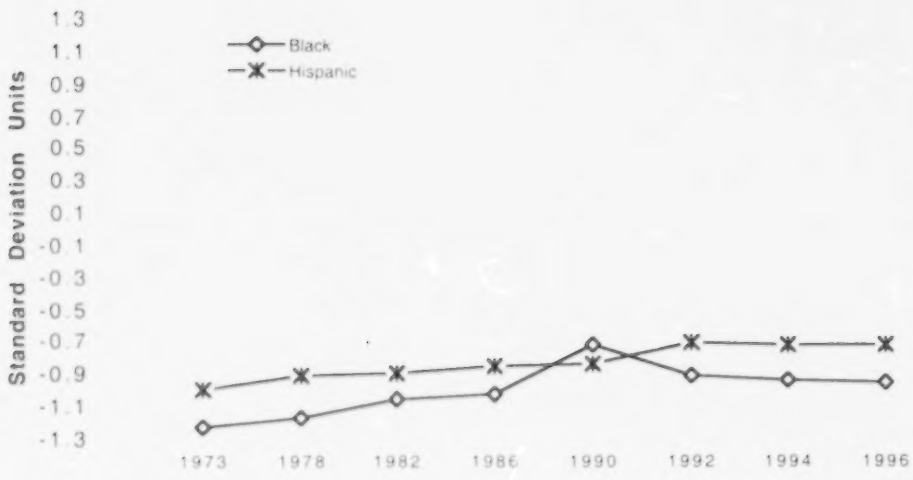

Figure 3. Trends in Racial/Ethnic Group Differences: NAEP Mathematics - 17-year-olds.

Note: The performance of White students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.

Sources: "Weighted Means, Standard Deviations, and Percentiles of Mathematics Distributions with Jackknifed Standard Errors" (Princeton, NJ: Educational Testing Service, undated and unpublished tabulations); John Mazzeo, Educational Testing Service, personal communication (September 1997); Campbell, Voelkl and Donahue (1997)

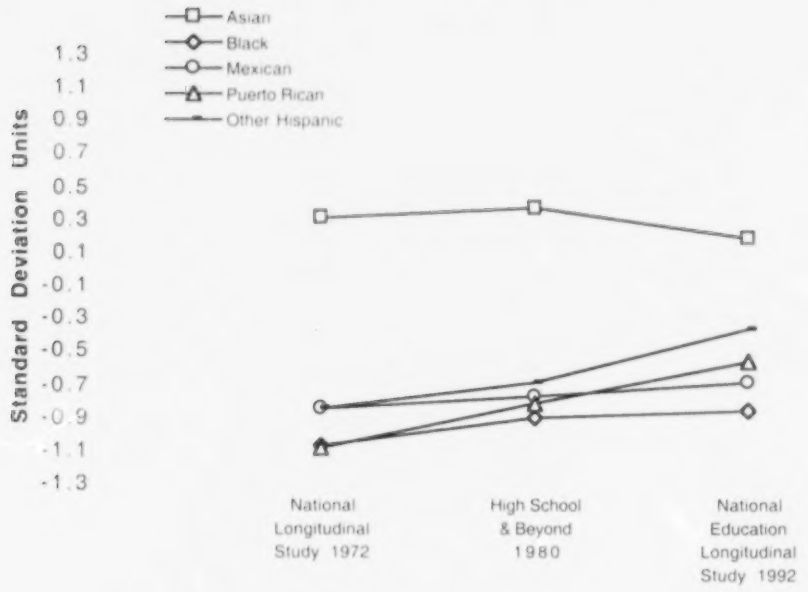

Figure 4. Trends in Racial/Ethnic Group Differences: NLS, HSandB, and NELS High School Seniors.

Note: The performance of White students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse.

Source: U.S. Department of Education, NCES (1995).

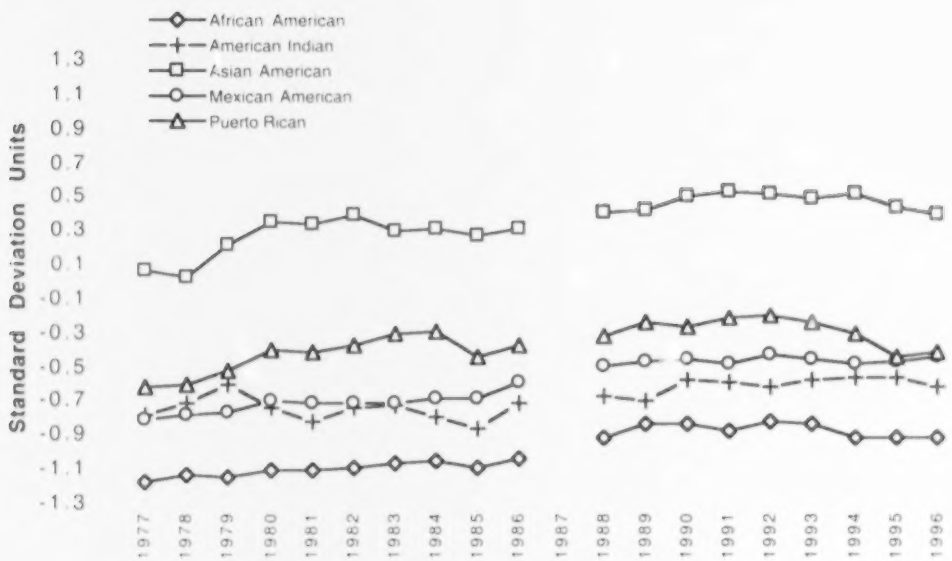

Figure 5. Trends in Racial/Ethnic Group Differences: ACT Mathematics - High School Seniors.

Note: The performance of White students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse. Data were not available for 1987.

Sources: "ACT Math Scores and ACT Score Means and SDs for Successive Years of ACT-Tested College-Bound Seniors 10% National Sample" (Iowa City: American College Testing, undated and unpublished tabulations); James Maxey, American College Testing, personal communications, (August-September, 1997).

Figure 6. Trends in Racial/Ethnic Group Differences SAT 1 Mathematics - High School Seniors.

Note: The performance of White students for each year is taken to be the zero point of the scale. Groups above the zero line perform better, groups below perform worse. Data were not available for 1986.

Source: College Entrance Examination Board (1972-1997).

A bicycle ride (1946)

I awakened early, jumped out of bed and had a quick breakfast. My friend, Mary Quant, was coming to our house at nine o'clock as we were going for a long bicycle ride together.

It was lovely morning. White fleecy clouds floated in the clear blue sky and the sun was shining. As we cycled over Castlemore bridge we could hear the babble of the clear stream beneath us. Away to our right we could see the brilliant flowers in Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

A day in the bog (1947)

I awakened early and jumped out of bed. I wanted to be ready at nine o'clock when my friend, Sadie, was coming to our house. Daddy said he would take us with him to the bog if the day was good.

It was lovely morning. White fleecy clouds floated in the clear blue sky. As we were going over Castlemore bridge in the horse and cart, we could hear the babble of the clear stream beneath us. Away to our right we could see the brilliant flowers in Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

A bus tour (1948)

I awakened early and sprang out of bed. I wanted to be ready in good time for our bus tour from the school. My friend, Nora Greene, as going to call for me at half-past eight as the tour was starting at nine.

It was lovely morning. White fleecy clouds floated in the clear blue sky and the sun was shining. As we drive over Castlemore bridge we could hear the babble of the clear stream beneath us. From the bus window we could see Mrs. Casey's garden. Early summer roses grew all over the pergola which stood in the middle of the garden.

Figure 7

Appendix A

Comparability of the TIMSS Population 3 Coverage Index ("SECI"), and Data on Dropouts in the United States

Keith Rust

Westat

Internal Memorandum

November 21, 1997

The preliminary tables produced by the TIMSS International Study Center for the TIMSS Population 3 results show a statistic known as "SECI" (for Secondary Coverage Index, but likely to be renamed in the final report). This statistic attempts to measure the size of the student population covered by the TIMSS sample, relative to the size of a typical single year age cohort around the age at which most students complete school.

In the draft report this figure is given as 65 percent for the United States. The NCES publication "Dropout Rates in the United States: 1995" shows the "status dropout rate" for the United States for 1995 as 12.0 percent. The same publication shows that of the grade 8 class of 1987-88, from spring of 1988 to spring of 1992, 10.8 percent of students dropped out.

Thus, at first glance the data on coverage produced by TIMSS and the data on dropouts published by NCES appear inconsistent. The purpose of this note is to examine the basis of the two sets of figures, to see if the

apparent discrepancy is explainable. First I will describe what is in the SECI, and then review something of the concepts used in measuring dropouts, before putting these together to see what inconsistency remains.

The calculation of SECI

The numerator of the SECI is calculated solely from the TIMSS student sampling data. It is the sum of the final sampling weights of all students assessed in TIMSS Population 3.

Thus, the SECI is an (unbiased) sample estimate of the number of students who:

a. were enrolled in grade 12 in 1994-95

b. were still enrolled in April 1995, near the end of the school year

c. were deemed assessable in TIMSS by school personnel

5. were in schools included in the TIMSS sampling frame – essentially schools on CCD and PSS in about 1993

School and student nonresponse in TIMSS did not systematically affect the SECI numerator. School and student nonresponse adjustments were applied to the assessed students' sampling base weights. There was no poststratification of any kind in the TIMSS weighting. Note that NAEP typically excludes about 3 to 4 percent of twelfth graders. The rate for TIMSS is likely to be similar, but we do not have a figure yet.

For the US, the TIMSS statisticians at Statistics Canada inform us that the numerator for the SECI is 2,346,705. Using the jackknife, Statistics Canada has calculated the standard error for this figure as 132,992.67, giving a 95% confidence interval of 2,086,047 to 2,607,368.

NAEP sampling procedures, time of assessment, and exclusion criteria are similar to TIMSS (NAEP takes place somewhat earlier in the year – January through March compared to April for TIMSS). Hence one would expect the size of the population to be rather similar between the two studies. NAEP does poststratify its estimates, however. For 1996, the NAEP estimates (from different parts of the sample) range from 2,464,876 to 2,593,809 before poststratification. The poststratification total is 2,379,702. Similar results are found in 1994 NAEP. Thus the NAEP data are quite consistent with TIMSS, before and after poststratification.

The fact that poststratification makes little difference to the NAEP results implies that any lack of coverage of the school sampling frame is not a significant source of population loss. This is reinforced for TIMSS by the fact that the grade 8 estimate of enrollment is much higher than grade 12, using the school sampling frames.

One would expect that most students who are in the TIMSS population would graduate with regular diplomas in 1995. It seems likely that very many students who are not going to graduate would have withdrawn from school by April. Given that many of the students excluded from TIMSS would also have graduated then, one would expect the number of graduates in 1995 to about equal the TIMSS population, and perhaps even exceed it slightly. The Statistical Abstract of the United States for 1996 quotes NCES data as showing that there were 2,505,000 graduates for the school year ending June 1995. Again, this seems quite consistent with the TIMSS SECI numerator, especially when keeping sampling error in mind.

The denominator of the SECI is defined as the number of persons in the US aged 15 to 19 in 1995 divided by 5. This definition was adopted in recognition of the fact that, across TIMSS countries, there is variation within countries as to the age of students when they complete school, depending upon the track they are in, and how much grade retention they have experienced.

For the US, the population aged 15 to 19 is 18,065,000 (Statistical Abstract of the United States, 1996). This gives a denominator of 3,613,000. The number of 18 year olds (as of July 1) is 3,506,000, and the number of 17 year-olds is 3,597,000. Thus it is likely that the SECI denominator for the US slightly overstates the appropriate cohort size for comparison. If one uses the number of 18 year-olds in the denominator, the SECI becomes 66.9%.

Taking into account the sampling error in the numerator, the standard error of the SECI is 3.7%, giving a 95 percent confidence interval of 57.5% to 72.2% (or, using the number of 18 year-olds in the denominator, 59.5% to 74.4%).

Thus, when one considers the nature of the population reflected by the TIMSS SECI and compares this with the NAEP data and published statistics on the number of graduates *receiving a diploma*, quite a consistent story emerges.

Statistics on Dropouts

"Dropout rates in the United States: 1995" shows a status dropout rate for 1995 of 12.0 percent. This means that of the 32.4 million persons aged 16 to 24 in October 1995, 3.9 million of them did not have a high school diploma or equivalent, and were not enrolled in school.

There are several things to note about this statistic. First, graduates include people with a GED or equivalent, who in fact did not obtain a regular high school diploma. Second, it is measured in October, at the beginning of the school year. Any students who drop out during the school year do not count as dropouts in this statistic. Third, I would estimate that approximately 25 percent of the people in the denominator are enrolled in school, and as such are by definition, not dropouts. By changing the definition of the population group covered to persons aged 13 to 24, this rate would reduce to 9%. By defining the group as those aged 18 to 24, it would probably increase to about 15%. It measures the proportion of the population that has already dropped out, and as such understates the number who will eventually drop out.

The same publication shows the event dropout rate for October 1995 as 5.7%. Roughly speaking this means that, on average across grades 10 through 12, 5.7 percent of the students enrolled in October 1994 were neither enrolled nor graduates in October 1995. Thus again this statistic does not reflect any incidences of students dropping out during the year and returning the following year – such students are not classified as dropouts. Also it is an average of three grades. Cumulating this rate over three years (and thus assuming 100% enrollment at the beginning of grade 10), one finds that 16.1 percent of students drop out.

These status and event statistics are obtained from the Current Population Survey. This survey uses self reporting and proxy reporting to obtain the information as to whether an individual is enrolled, was enrolled last year, and has or does not have a high school diploma. Although I have no evidence about this, it seems likely that such self-reporting might tend to lead to an understatement of the extent of dropouts.

The same publication shows cohort dropout rates from the NELS study. From the eighth grade class of 1987-88, of the students enrolled in grade 8 in the spring of 1988, 10.2 percent were no longer enrolled by the spring of 1992. But many of those who were still enrolled will not eventually graduate, at least not with a regular diploma. Also NELS data is based on following up the original sample and it seems likely that attrition bias for dropout data could be very substantial (again, I have no evidence).

These statistics are measures of the extent of dropping out. But based on the way the SECI is defined, one would expect a closer correspondence with estimates of the rate of graduation with a regular diploma. Table 14 of the 1995 dropout publication shows that, of those 18 to 24 year-olds not currently enrolled in school in October 1995, 77.9 percent had an equivalent qualification (and so are classified as graduates when calculating status dropout rates, but would not appear in the SECI numerator). Given that between October and April the pool of persons aged 18 to 24 will increase due to dropouts, but the number of regular graduates will hardly change, if this statistic were estimated in April (the time of the TIMSS assessment) it would no doubt be considerably lower, and consistent with the SECI once one accounts for TIMSS exclusions.

Reconciling the SECI and Dropout Statistics

It can be seen that what appears initially to be a large discrepancy between the TIMSS SECI and published dropout statistics can essentially be explained by definitional differences. However, much of the explanation relies on the assumption that many students who are enrolled in grade 12 in October have left school by April/May. Is there any evidence to back this up?

The first and most convincing evidence is the statistic cited above about the number of diplomas awarded in 1995 (2,505,000). This figure is inconsistent with the idea that, say, only 12 percent of the 3.5 million students in a cohort drop out (as this would imply over 3 million graduates).

There are two other pieces of evidence from the TIMSS study itself. Because of the complicated sampling scheme, involving different sample targets and different sampling rates for students with different course taking patterns, Westat approached sample schools several months before the testing and asked them approximately how many students were enrolled. About half the schools provided the necessary information. When actual student lists for sampling were drawn up, on average across the 100 or so schools, only about 80 percent as many students were identified as were indicated in the initial survey.

From these lists that were provided, samples of students were drawn. Some of these students were coded as no longer enrolled. Statistics Canada tells us that the weighted estimate of the number of such students is 88,000 or about 3.7 percent of the population. Yet we requested these lists of students only a few weeks prior to the testing (although no doubt in some cases we received lists that reflected the fall enrollment). This evidence is too sketchy to be used to quantify the extent to which students drop out during grade 12 (perhaps to graduate over the summer, return next year, or later earn a GED), but this does provide evidence that the numbers are probably substantial.

In conclusion, close examination of the concept of the TIMSS SEC1, and the way it is calculated, contrasted with the methods used to measure

dropout rates for the US, shows that in fact there is no great inconsistency. The directly comparable data available suggests that the TIMSS SECI is an accurate measure, at least when one considers the effects of sampling error, and variations in the size of single year age cohorts.

---

1 This is a grim scenario that continues to be taken for granted and hawked by many observers, commentators, commissions, academics and policy-makers. There are ,however, opposite arguments based on substantial evidence that our educational problems have been miss-portrayed, vastly over generalized, and are neither universal, nor nation threatening. Nonetheless, good news about the public schools is either ignored or gets short shrift (Berliner, 1993; Bracey, 1992; 1993; Madaus, 1995; Tanner 1993) For a full discussion of why is it that the proclamation of generalized bad news eclipses the good, see Madaus (1995).

2 For a more complete listing of the purported benefits of high stakes "authentic" assessments, see Madaus (1993).

3 The average for White 13-year-olds on the NAEP scale was 281 scale score points; for Black 17-year-olds it was 286.

4 This fact was recognized in the early 60s when intervention programs like Head Start and Title 1 were initially justified on the basis that an "achievement gap" existed between disadvantaged and other children. This gap was defined in terms of standardized-test performance levels. (see Biemiller, A.J. 1965.)

Anthony Celebrezze, Secretary of Health, Education, and Welfare testifying before a congressional hearing committee arguing for the passage of the ESE Act of 1965 cited standardized test results to justify enactment. He pointed out:

You will find that by the end of the third year[grade] this student [in central Harlem in New York City] is approximately 1.2 grades behind the national average and 1.1 grades behind the New York City average. By the time he gets to the sixth grade, he is 2.1 grades below the national average and two grades below the New York average. And by the time he gets to the eighth grade, he is 2 1/2 grades below the national average and approximately 2 grades below the New York average. . . . The students continue to get further and further behind in terms of standardized test norms. . . . (Celebrezze, 1965, p. 89)

5 For example, consider grade 12 where only 2% of students are classified at the advanced level. Nonetheless, about 50% of all seniors sit for either the SAT or ACT each

year. About 16% of the 50% score higher than one standard deviation above the mean; this constitutes about 8% of all seniors, four times as high as the 1994 NAEP advanced category percentage. In other words, being in the top sixteen percent of students on the ACT or SAT math means that large numbers of these seniors would not reach the NAEP "advanced" attainment level. Does this make intuitive sense; are the NAEP math achievement levels defensible, particularly given that the SAT and ACT are important in the students' lives while the NAEP is a "drop from the sky" event without any personal consequences?

[6] For a development of these principles see Madaus, (1988)

[7] You cannot measure either an electron's position or velocity without distorting one or the other. How this principle plays out for high stakes testing is described in the three remaining principles.

[8] The validity of this assertion was confirmed in the late 80s by a West Virginia physician, Dr. John Cannell. Surprised and curious about the above average scores obtained by some local students on nationally normed achievement tests, Cannell collected standardized test results from states and school districts across the country. He found that most states and districts were reporting above average scores. When subsequent study of what came to be called the "Lake Wobegon" effect was undertaken, one of the explanations advanced for the almost universally above average results was that schools routinely taught directly to the test, and even to specific, known test questions (Cannell, 1987, 1989).

[9] Thus, back in the 1930s, writing about the New York State Regents examination, which determined who would receive a Regent's diploma, Spaulding (1938) noted that teachers felt they had to abandon locally-developed curriculum guides in favor of the curriculum defined by the Regents examinations, if they were to assist their students over the important examination hurdle. A recent study in Israel based on observations and interviews over time, concluded that the introduction by the Ministry of Education of three different tests all perceived to have high-stakes, narrowed the process of education, "making it merely instrumental and unmeaningful" (Shohamy, 1993, p17).

[10] A clear example of this surfaced during the Minimum Competency Test hearings (NIE, 1981), sponsored by the National Institutes of Education. A principle from New York explained how high-stakes multiple-choice tests had affected reading instruction in her school. The tests not only dictated the content focus of instruction, but also the form that instruction took. The principal told how her students practiced "reading" by reading dozens of little paragraphs and answering related multiple choice

questions. Further, when a section on synonyms and antonyms was dropped from the test, the practice materials on synonyms and antonyms were dropped from the teacher's arsenal of instructional techniques.

[11] Also see Hamilton, L.S., Nussbaum, E.M., and Snow, R.E., "Interview Procedures for Validating Science Assessments," *Applied Measurement in Education*, vol. 10(2), 1997, pp. 181-200.

[12] The study had two parts. First a closed-ended questionnaire was mailed directly to 4950 public school teachers in grades 4 to 12. This sample was stratified by three variables: content area (math and science, using two separate parallel questionnaires); urbanicity (urban, suburban and rural school districts; and school level (elementary, middles/junior and secondary). From two mailouts in January and February 1991, 2229 teachers responded for a response rate of 45%, which was considered to be adequate given the length of the survey (7 pages) and the average amount of time for completion (1 hour). The teachers who responded were similar to the U.S. population of teachers as indicated by NCES 1990 data. Second were six case studies of teachers and administrators in six urban districts with large minority enrollments. For technical details and a full report of the finds see Madaus, West, Harmon, Lomax and Viator, 1992, and West, M. M., & Viator, K. A. (1992).

[13] In the 18th and 19th centuries in Britain pioneers of examining at that time believed that self-interest was the main motive for study and, since study involved drudgery, it was necessary to link important rewards or sanctions to learning (Madaus & Kellaghan, 1991a;1991b). Adam Smith (1990) in the 18th century expressed the need for extrinsic rewards linked to examinations when he wrote

> The public can encourage the acquisition of those most essential parts of education, by giving small premiums, and little badges of distinction, to the children of the common people who excel in them.
>
> The public can impose upon almost the whole body of the people the necessity of acquiring those most essential parts of education, by obliging every man to undergo an examination or probation in them before he can obtain the freedom in any corporation, or be allowed to set up any trade either in a village or town corporate. (p.384)

[14] See Webb, Covington and Guthrie 1993 for a further discussion

[15] These positions are mirrored by advocates of testing who emphasize an instrumental, qualification-gathering aspect of education (see Brooke & Oxenham, 1984; Dore, 1976) and a view of instruction as being driven by measurement (Measurement Driven Instruction), now often described as outcome-based education.

with its three ingredients: a clear concept of educational goals, or in current parlance, standards; a test that measures the goals; and high stakes associated with test results to act as a driving force (Popham, 1983, 1987; Popham, Cruse, Rankin, Sandifer, & Williams, 1985).

[16] See (Dweck & Leggett, 1988).

[17] See (Deci & Ryan, 1985; Deci, Spiegel, Ryan, Koestner, & Kauffman, 1982; Deci, Vallerand, Peletier, & Ryan, 1991).

[18] The "authentic" assessment movement emerged in the late 1980s (Mitchell, 1992, Wiggins, 1989). Underpinning the movement is the belief that student learning and progress are best assessed by tasks that require active engagement, such as producing extended responses or some other tangible product that can be evaluated on its merits, investigating complex problems, generating material for portfolios, performing exhibitions, or carrying out experiments, rather than by having students select an answer from several alternatives.

[19] Data on how the use of performance assessment might impact different groups over a long time frame simply do not exist in the UK or elsewhere

[20] Thomas, Madaus, Raczek and Smees collected both teacher assessments (TA) and standard task (ST) assessments for 17,718 in 590 schools in one large Local Education Authority across 9 attainment targets. In English these topics were reading, writing, spelling and handwriting; in mathematics the topics were number operations, data handling: collecting, recording and processing data and data handling:probabilities; and in science the topics were types and uses of materials and earth and atmosphere. The STs are examples of what people in the "authentic" assessment movement in this country call for. The TAs are just that -- teacher judgments about student attainment on the same areas of the national curriculum measured by the STs. For further details of the study see Thomas, S., Madaus, G. F., & Razcek, A. (1998).

[21] (Shepard & Smith, 1988; see Shepard & Smith for a discussion of validity issues surrounding the use of "readiness" tests for entrance or retention).

[22] Kreitzer, Madaus and Haney calculated what is sometimes called "attrition rate." Attrition rates are calculated by subtracting the graduation rate from 100 percent. The graduation rates were calculated by the Department of Education by dividing the number of public school graduates by the ninth-grade enrollment four years earlier. The rates were adjusted by DOE for migration and students who are unclassified by grade. For details see Kreitzer, Madaus, Haney (1989).

[23] Data from the 1988 and 1990 National Educational Longitudinal Surveys (NELS) on students who were required to pass one or more MCTs in eighth grade in 1988 was examined in order to find out if such students were more likely to have dropped out of school by tenth grade than students who did not have to meet such a requirement.

[24] Fassold identified a student as a TAAS drop out if (A) he/she failed an exam after which he/she missed the remaining exams before his/her class's scheduled graduation; and conjunctively (B), did not drop out for one of 18 specified reasons on an Dropout-Exit-Reason Code used by the state, and the student did not defer from taking the test. Some of the reason for dropping out included job, military, pregnancy, poor attendance, and age

[25] The SES is dependent solely on participation in the school lunch program. Free meals reduced lunch and not eligible.

[26] See Table B5 in Appendix B of the report entitled Mathematics and Science Achievement in the Final Year of Secondary School

[27] (One of the 17 Commissioners endorsing this recommendation was Bill Clinton, then Governor, State of Arkansas.

[28] (Zimmerman, Bandura, & Martinez-Pons, 1992)